Sir David Cox's Statistical Philosophy and Its Relevance to Today's Statistical Controversies

15 July 1924 – 18 January 2022

Deborah G. Mayo

Virginia Tech *August 8, 2023*

JSM 2023, Toronto

Cox is celebrated for his pioneering work for which he received many awards…

- International Prize in Statistics in 2016.

- Copley Medal, the Royal Society’s highest award, 2010.

- Knighted in 1985.

- Gold Medal for Cancer Research for the "Proportional Hazard Regression Model" in 1990.

Importance of statistical foundations

Alongside his contributions that have transformed applied statistics, Cox focused on the foundations of statistical inference

"Without some systematic structure statistical methods for the analysis of data become a collection of tricks…" (2006, xiii)

How does a philosopher of statistics wind up working with a giant in statistics?

In late summer 2003 (nearly 20 years ago today), I (boldly) invited him to be in a session I was forming on philosophy of statistics

[Optimality: The Second Erich L. Lehmann Symposium, May 19–22, 2004; Rice University, Texas]

Discussions over the next several years:

- "Frequentist Statistics as a Theory of Inductive Inference" (Mayo and Cox 2006)

- "Objectivity and Conditionality in Frequentist Inference" (Cox and Mayo 2010)

“A Statistical Scientist Meets a Philosopher of Science”

Sir David Cox: “…we both think foundations of statistical inference are important; why do you think that is?”

Mayo: “…in statistics …we invariably cross into philosophical questions about empirical knowledge and inductive inference.” (Cox and Mayo 2011)

Key Feature of Cox's Statistical Philosophy

- We need to **calibrate** methods: how they would behave in (actual or hypothetical) repeated sampling

  "any proposed method of analysis that in repeated application would mostly give misleading answers is fatally flawed" (Cox 2006, 198)

  Weak repeated sampling principle

Two philosophical questions (still controversial):

1. How can the frequentist calibration be used as an evidential or inferential assessment (epistemological use)?

2. How can we ensure: “that the hypothetical long run used in calibration is relevant to the specific data” (Cox 2006, 198)

Mayo & Cox 2006
"Frequentist statistics as a theory of inductive inference"

"…we begin with the core elements of significance testing in a version very strongly related to but in some respect different from both Fisherian and Neyman-Pearson approaches…" (2006, 80)

R. A. Fisher

J. Neyman

E. Pearson

Simple statistical significance tests

"…to test the conformity of the particular data under analysis with H_0 in some respect:

[We find a] **test statistic *T***, such that

- the larger the value of T the more inconsistent are the data with H_0;

…the p-value $Pr(T \geq t_{obs}; H_0)$" (81)

Testing rationales: *Inductive behavior vs inductive inference:*

Small p-values *are taken as inconsistency with H_0* (in the direction being probed)

- *Behavioristic*: To follow a rule with low error rates.
- *Evidential*: The specific data set provides evidence of discrepancy from H_0. *Why?*

(FEV) Frequentist Principle of Evidence: Mayo and Cox (2006)

"**FEV(i):** *y* is (strong) evidence against H_0 i.e., (strong) evidence of discrepancy from H_0 if and only if [were H_0 correct] then, with high probability [(1-p)] this would have resulted in a less discordant result than [observed in] *y*" (82)

Inferential interpretation

“a statistical version of the valid form…*modus tollens* [falsification]

[here] probability arises …to characterize the ‘riskiness’ or probativeness or severity of the tests to which hypotheses are put…reminiscent of the philosophy of Karl Popper” (82)

FEV (ii) for interpreting non-statistically significant results, avoiding classic fallacies

Move away from a binary "significant / not significant" report

- One way: consider testing several discrepancies from a reference hypothesis H_0 (i.e., several H_i's) and reporting how well or poorly warranted they are by the data

This gets beyond long-standing and recent problems leading some to suggest restricting statistical significance tests to quality control decisions (behavioristic use)

Analogy with measuring devices

"The significance test is a measuring device for accordance with a specified hypothesis calibrated …by its performance in repeated applications,

…we employ the performance features to make inferences about aspects of the particular thing that is measured, aspects that the measuring tool is appropriately capable of revealing." (84)

Relevance

"in ensuring that the … calibration is relevant to the specific data under analysis, often taking due account of how the data were obtained."
(Cox 2006, 198)

Selection effects

Selectively reporting effects, optional stopping, trying and trying again etc. may result in failing the weak repeated sampling principle.

Selection effects

"the probability of finding some such discordance or other may be high even under the null. Thus, following FEV(i), we would not have genuine evidence of discordance with the null, and unless the p-value is modified appropriately, the inference would be misleading." (92)

Not all cases with selection or multiplicity call for adjustments

Searching a full database for a DNA match:

- "The probability is high we would not obtain a match with person *i*, if *i* were not the criminal;
- By **FEV**, the match is good evidence that *i* is the criminal.
- A non-match virtually excludes the person." (94)

Calibration is the source of objectivity: "Objectivity and Conditionality in Frequentist Inference" (Cox & Mayo 2010):

> "Frequentist methods achieve an objective connection to hypotheses about the data-generating process by being constrained and calibrated by the method's error probabilities" (Cox and Mayo 2010, 277)

“Objectivity and Conditionality in Frequentist Inference” (Cox & Mayo 2010)

When Cox proposed this second paper include “conditioning", I hesitated….

Good long-run performance isn't enough

Cox's (1958) "two measuring instruments" ("weighing machine" example): toss a fair coin to test the mean of a Normal distribution with either a very large, or a small, variance (both known); and you know which was used.

Good long-run performance isn't enough

As Cox (1958) argued, once you know the precision of the tool used

"the p-value assessment should be made conditional on the experiment actually run." (Cox and Mayo 2010, 296)

Weak conditionality principle

- The “weighing machine” example caused “a subtle earthquake” in statistical foundations. Why?

- The “best” test on long-run optimality grounds is the unconditional test.

Does "conditioning for relevance" lead to the unique case?

"[Many argue a frequentist] "is faced with the following dilemma: either to deny the appropriateness of conditioning on the precision of the tool chosen by the toss of a coin, or else to embrace [a] principle* … that frequentist sampling distributions are irrelevant to inference [postdata]. This is a false dilemma. Conditioning is warranted in achieving objective frequentist goals." (Cox & Mayo 2010, 298)

*Strong Likelihood Principle

- The (2010) paper with Cox was the basis for my later article in *Statistical Science* debunking the argument that appears to lead to the dilemma: Mayo (2014)

“Principles of Statistical Inference” (2006)

“The object of the present book is to [describe the main] controversies over more foundational issues that have rumbled on …for more than 200 years”.

- Cox (1958) “Some problems connected with statistical inference”
- Cox (2020) “Statistical significance”

“His writing on statistical significance and *p*-values seemed to need repeating for each new generation”

(Heather Battey and Nancy Reid 2022, IMF obituary for Sir David Cox)

Cox advanced a great many other foundational issues I have not discussed:

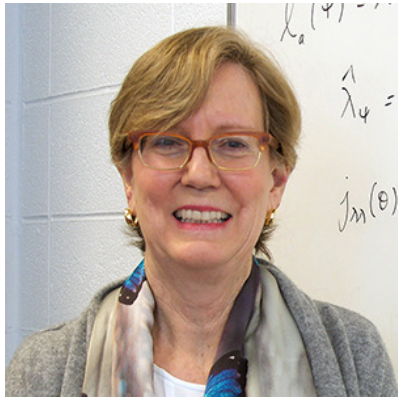
- relating statistical and scientific inference;
- testing assumptions of models;
- rival paradigms for interpreting probability, frequentist and Bayesian

David R. Cox Foundations of Statistics Award

Wednesday, August 9: 10:30-12:20

Metro Toronto Convention Centre Room: CC-701A

Chair: Ron Wasserstein (ASA)



Inaugural recipient: Nancy Reid, University of Toronto

“The Importance of Foundations in Statistical Science”

References

- Battey, H. and Reid, N. (2022). "Obituary: David Cox 1924-2022". Institute of Mathematical Statistics homepage, April 1, 2022.
- Cox, D.R. (1958). Some problems connected with statistical inference. *Ann. Math. Statis*
- Cox, D. R. (2006). *Principles of Statistical Inference*, CUP.
- Cox, D. R. (2018) In gentle praise of significance tests. 2018 RSS annual meeting. (Link to Youtube video here.)
- Cox, D. R. & Mayo D. G. (2010). Objectivity and Conditionality in Frequentist Inference. In Mayo & Spanos (eds), *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability and the Objectivity and Rationality of Science*, pp. 276-304. CUP
- Cox, D. R. and Mayo, D. G. (2011). A Statistical Scientist Meets a Philosopher of Science: A Conversation between Sir David Cox and Deborah Mayo. *Rationality, Markets and Morals* (RMM) 2, 103–14.
- Mayo, D. G. (1996). *Error and the Growth of Experimental Knowledge*, Chicago: Chicago University Press. (1998 Lakatos Prize)

- Mayo, D. G. (2014). "On the Birnbaum Argument for the Strong Likelihood Principle," (with discussion) *Statistical Science* 29(2) pp. 227-239, 261-266.
- Mayo, D. G. (2018). *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*, Cambridge: Cambridge University Press.
- Mayo, D. G. (2022). A Remembrance of Sir David Cox: '"In celebrating Cox's immense contributions, we should recognise how much there is yet to learn from him". *Significance*. (April 2022: 36). [Link to all 4 remembrances.]
- Mayo, D.G. and Cox, D. R. (2006) "Frequentist Statistics as a Theory of Inductive Inference," *Optimality: The Second Erich L. Lehmann Symposium* (ed. J. Rojo), Lecture Notes-Monograph series, IMS, Vol. 49: 77-97. [Reprincted in Mayo and Spanos 2010.
- Reid, N. and Cox, D. R. (2015). On Some Principles of Statistical Inference. *International Statistical Review* 83(2): 293-308.